DOCUMENT RESUME

ED 183 613 TM 800 021

AUTHOR McLarty, Joyce R.
TITLE Establishing Passing Scores.
PUB DATE Nov 79
NOTE 36p.: Paper presented at the Annual Meeting of the California Educational Research Association (San Diego, CA, November 15-16, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; *Cutting Scores; Decision Making; *Evaluation Criteria; *Mastery Tests; *Psychometrics; *Social Values; Test Interpretation

ABSTRACT

The problem of establishing appropriate passing scores is one of evaluation rather than estimation and not amenable to exact solution. It must therefore be approached by (1) identifying criteria for judging the acceptability of the passing score, (2) collecting the data appropriate to assessing each relevant criterion, and (3) judging how well the criteria are met within successively restricted score ranges until convergence on the best possible passing point is reached. Two major features of passing scores are identified: the psychometric (score) components and the decision (passing) components. A review of the literature suggests that criteria appropriate to the psychometric components include the validity, reliability, and accuracy of the test within the range of concern. Criteria appropriate to the decision components include ethics and morality, acceptability, revisability, socio-political awareness, educational benefit, efficiency, and explainability. Each criterion is discussed in terms of its relevance to typical situations and of the types of data necessary to assess it. The problems of multiple test forms, post remediation reassessments, subgroups of examinees with special needs, and the assessment of multiple content areas (as in high school minimum competency testing) are discussed. An example of a process by which an appropriate passing score can be established is offered. (Author/JKS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED183613

Establishing Passing Scores

Paper presented at the
annual meeting
of the
California Educational Research Association

San Deigo, California
November 15-16, 1979

Joyce R. McLarty

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Joyce R. McLarty
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800021

ANNUAL MEETING OF CALIFORNIA EDUCATIONAL RESEARCH ASSOCIATION

Bahia Hotel, San Diego, California

November 15-16, 1979

Establishing Passing Scores

Title of Paper or Symposium

Author(s) Joyce R. McLarty

Position(s) Student

Institutional Affiliation(s) U.C.L.A.

Mailing Address 3890 E. Lucas Ct. Simi Valley, Ca. 93063

Street City State ZIP

Previous authors have failed to solve the problem of establishing appropriate passing scores because they have assumed that the problem was one of estimation. This paper contends that the problem is one of evaluation rather than estimation and not amenable to exact solution. It must therefore be approached by 1)identifying criteria for judging the acceptability of the passing score, 2)collecting data appropriate to assessing each relevant criterion, and 3)judging how well the criteria are met within successively restricted score ranges until convergence on the best possible passing point is reached.

Two major features of passing scores are identified: the psychometric (score) components and the decision (passing) components. A review of the literature suggests that criteria appropriate to the psychometric components include the validity, reliability, and accuracy of the test within the range of concern. Criteria appropriate to the decision components include ethics and morality, acceptability, revisability, socio-political awareness, educational benefit, efficiency, and explainability. Each criterion is discussed in terms of its relevance to typical situations in which passing scores are established and of the types of data necessary to assess it. The problems of multiple test forms, post remediation reassessments, subgroups of examinees with special needs, and the assessment of multiple content areas (as in high school minimum competency testing) are discussed. An example of a process by which an appropriate passing score can be established is offered.

Copies of the paper(s) ~~are~~ are not available from the author(s) upon request.

A discussion of how to establish appropriate passing scores may well begin with the question "appropriate for what?" Passing scores are those test scores which are used operationally to divide the examinee population into two subgroups, labeled "passing" and "failing." The reasons for undertaking this division vary, but fall into five general categories: description, motivation, certification, selection, and placement. Most passing score systems can be classified in at least two of these categories, and many into three or more. Description is perhaps the weakest use of pass/fail systems. In such a use, the instructor may simply wish to know how many individuals have performed up to a pre-determined standard in order to revise the instructional program. Obviously, purely descriptive uses not associated with some external standard do not require the use of a passing score since the score distribution provides more information and the setting of an after-the-fact passing score for descriptive use is an entirely circular process.

Motivational effects of pass/fail systems depend for the most part on the other uses which are made of the information. If no use is made of the score, motivational effects will be nil. If the score is used, but not the pass/fail designation (as for example where quiz scores are to be added in determining the final grade), the "fail" designation may motivate students receiving it to increased efforts. If the scores are used for certification, selection, or classification, motivational effects will depend on the specific situation.

Certification refers to that usage of passing scores which permits those passing to exit a given program. Those who fail are typically retained or given remedial work. Selection usage allows those passing to enter a given program, with the primary concern usually being their anticipated performance in it. Little concern is generally paid to those who fail in a selection system. Finally, classification systems route individuals who pass to one activity and those who fail to another, with post examination concern for both groups. The setting of an appropriate passing score will require consideration of which usage(s) are intended.

These aspects of the passing score, which relate to its effects, identify it as the numeric representation of a decision system. Decision systems are, in general, reactive. The passing score decision system is affected by other social systems, and in turn affects them. For this reason, the passing score cannot be established without consideration of the context. Efforts to treat establishment of passing scores as estimation problems are doomed to failure because they do not recognize the ongoing, reactive effects of such systems.

Unlike some other decision systems, however, passing score systems are psychometrically based. Passing scores are indeed test scores and as such are tied to the reliability, validity, accuracy, and psychometric scale properties of the test instrument on which they are based. For this reason, efforts to treat establishment of passing scores in terms of strictly socio-political considerations are likely to be inappropriate, or even unethical.

Because of the duel nature of passing score systems, the establishment of appropriate passing scores can be rather tricky. Scores which seem to function well in the decision-system context may be very different from those which appear appropriate psychometrically. Indeed, the very concept of the "right" passing score is inappropriate since there may be none, or more than one. Instead, efforts need to be directed toward selecting the "most appropriate" passing score, given the psychometric qualities of the instrument, the socio-political context, and the decision to be made.

Efforts to evaluate possible passing scores to determine which is the "most appropriate", like any evaluation efforts, are based on the identification of relevant criteria, the determination of the relative importance of the criteria, and the measuring of that which is to be evaluated (in this case the passing score) against the criteria. Information regarding how well a given passing score satisfies each relevant criterion, and how important each criterion is in the current situation, combines to provide an index of the desirability of that passing score. The most desirable may then be selected.

The literature suggests a number of criterion measures against which the desirability of a passing score may be judged. Although the wording and emphases differ, there appear to be ten general categories by which passing scores are determined to be adequate or inadequate (American Friends Service Committee, 1974; Berk, 1976; Conaway, 1976; Shepard, 1976; Dwyer, 1977; Hambleton, 1978; Jaeger, 1978; Popham, 1978; Scriven, 1978). Three of these categories contain criteria relevant to the psychometric aspects of passing scores. They are reliability, validity, and accuracy. The remaining seven relate to the decision aspects. They are ethics and morality, acceptability, revisability, socio-political awareness, educational benefit, efficiency, and explainability. Each category combines a number of related suggested criterion considerations.

Reliability, in its psychometric sense, concerns the degree of confidence one has that the test score is stable for a given level of examinee functioning under acceptable conditions; that as long as examinee competence is stable, so is the test score. Reliability is treated quite adequately in standard texts on psychometric measurement, and will not be discussed in detail here. It is important to note, however that the reliability of the test instrument is not the same as the reliability of the decision. Where a passing score is used to separate the examinee population into two groups, it is the reliability of the classification rather than the reliability of the instrument which is of primary concern. It is more important that the score be consistantly above or below the given point than that the actual score level be consistent. Obviously a highly reliable instrument will meet both demands. A less reliable one may still provide a reliable decision if the decision point is in a region of the score distribution where a score change of a few points will not change the decision for very many individuals. Operationally, this means spreading out the scores in the decision region, usually by increasing the number of items or observations.

Validity, psychometrically speaking, addresses the degree to which one is willing to believe that the test

score is sufficient evidence for inferring the actual or predicted level of a specified competence. Like reliability, validity is treated in texts on psychometrics. As is the case with reliability, the chief concern for passing score systems is not the validity of the instrument but the validity of the decision.

While the reliability of the decision can often be improved by increasing the number of items or observations, improvement of the validity of the decision often requires changing the nature of the instrument. In a criterion-related validation system, such as might be appropriate in a selections system of decision-making, improving the validity of the instrument usually means finding more kinds of competencies related to the criterion performance to assess. Since the main function of the instrument is prediction, so long as the competencies are related to the criterion and are not illegal to use (eg. race or religion), the problem is simply finding ways to measure more things.

In a domain referenced validation system, such as might be appropriate in a certification system of decision-making, the validity of the instrument is based on generating a domain of items covering the competencies taught and taking a sample of them to form the test. Where this provides inadequate validity for the decision, as evidenced by dissatisfaction with the qualifications of those passed or failed, one may be faced with the need to tailor the test so that more items which properly distinguish the two groups are included. If, for example, the passing point is located at a dense point in the score distribution, more items at about that level of difficulty would improve the separation. However writing a sample of items covering the content domain but remaining at that level of difficulty is not now within our capabilities and may never be. Given the usual limitations on testing time as well, the most appropriate move would generally be to restrict the domain itself by removing those content topics which are much easier or more difficult. A sample from the restricted domain then remains readily interpretable but is likely to contain enough items at the desired level of difficulty to provide for a valid decision. Note that the function of the test

to provide both a decision and a description is retained except that those topics which were at an unusuable level of difficulty, that is, which were irrelevant to the decision to be made, are excluded from both the instrument and the description.

Accuracy addresses concerns for the individual examinee. As with reliability and validity, the concern narrows from the full range of scores to that narrow band within which the decision becomes critical. It doesn't much matter to the decision if the error of measurement is gigantic, so long as it doesn't overlap the decision point. The same thing is true of the error of estimate in a criterion-based decision system. One would like to believe that the probability that an examinee is correctly classified, given a standard of comparison (domain performance, criterion variable, or level on a latent trait), is at least as good using the passing point as it would be using the teacher's opinion, or the examinee's self-rating, for example.

Taken together, the psychometric criterion categories attempt to ensure that the test score at or near the passing point provides an acceptable estimate of the competence or characteristic in question. As suggested above, difficulties in meeting these criteria are generally met by improving the test instrument itself, usually by tailoring it to the specific use.

The decision criterion categories are both less homogeneous and less specific than the psychometric criterion categories. While it may be possible to improve the reliability, validity, and accuracy of the test simulteneously, improvement with respect to one decision criterion will often lead to the decrement of another. In addition, while psychometric concerns can often be handled in consultation with expects on an in-house basis, decision concerns are generally public considerations over which the testing agent has little control. While changes in the test instrument can occasionally be used to mitigate difficulties with decision criteria, in most cases the focus is on the placement of the passing score itself.

Ethical and moral concerns are criteria which dictate that the decision process and its results be fair to examinees

and their families according to current standards. Even where the information would improve the accuracy of the decision, certain information is considered inappropriate for use (eg. race, sex, and religion). Ethical standards also dictate that handicapped individuals and those whose native tongue is not English receive special testing except where determination of their actual competence would be impossible by any alternative method (that is, it is not appropriate to test English ability in Spanish, but ability to read may be tested in braille).

Other ethical considerations suggest that a system of appeals be established within the decision system so that examinees and their families can protest any part of the process which they may feel is unfair. Finally, it would be considered ethically advisable to inform prospective examinees of the purpose, nature, and general form and content of the examination in order that they may be appropriately prepared for it.

Acceptability refers to the willingness of the social systems involved to accept the decision system embodied in the passing score, including its anticipated and actual effects on them. Some authors have suggested that the appropriate consideration is whether a similarly informed group of judges would have selected the same score. That approach assumes that there is some "right" score to be estimated, and that if different judges agree on that score one can be more confident of its correctness. Actually, it is less important that others would have selected the same score than that they be willing to live with it.

Popham (1978) suggests that preference data be examined in determining the acceptability of the passing score. Then if the context suggests that no passing score below 65, for example, is likely to be acceptable, action can be taken by either setting the score higher directly or by rescaling the test so that the higher number is appropriate.

Educational benefit concerns a knowledge that teachers do in fact "teach to the test". Although test validity is the primary concern (so that teachers "teach to" something which

is congruent with the curriculum), the motivational effects of the passing score itself must also be considered. If it is too high, the effect may be demoralization; if it is too low, lethargy.

An additional concern with respect to educational benefit is the need to "feed back" to the educational system those test results which might benefit the instructional program in general or the individuals who fail specifically. Although it will generally be unadvisable to create diagnostic tests for use in passing score systems, some information from the test may be useful to the remedial program

Efficiency is basically a fiscal category. In its broadest form, it asks "Why spend money on testing when the teacher could make just as good a decision?" That question generally deserves more attention than it receives. Additionally, the efficiency criteria suggest that information which is routinely collected for other purposes, and information which is available at little or no cost in time and money be considered for use in the system; that less costly methods replace more costly ones wherever possible. Finally, efficiency dictates consideration of the purposes for which the passing score is used in determining how many are to be passed or failed. For example, it may be inefficient to set a stringent passing score requiring a large number of individuals to repeat a course, where those who pass are routed to the next course in which they are immediately provided with a review which would bring all but the most severely deficient up to the original passing level.

Revisability acknowledges the reacttivity of the decision system. The more important the decision, in general, the more reactive the system is likely to be. As the system reacts to the score, the score itself may need to change to remain compatible. A continuing review of the test score and examinee placement distributions provides information which can be used to determine whether the appropriate passing score has changed.

An additional concern in this category relates to the original establishment of the passing score. The human mind cannot process the incredible number of complex variables which are involved in the passing score decision simultaneously.

It is therefore recommended that the process of setting the passing score make use of multiple passes through the data, each followed by feedback as to the effects of the previous decision. This itterative process allows for successive approximation to the final, "best" passing score.

Socio-political awareness is the "catch-all" category into which all of the remaining social system concerns are placed. It may include such diverse concerns as the political ambitions of school board members, the pet project of the PTA, and the academic performance of the captain of the football team. Some of these concerns relate to ethics, some to educational benefit, and all at least indirectly to acceptability. However, the key issue in classifying them here is that socio-political concerns are very specific to an individual situation, and are generally unresponsive to direct evidence of the appropriateness of the passing score system. That is, they tend to be emotional more than logical concerns. For example, parent concerns relating to the validity of the test for black students would not be directly related to the level at which the passing score was set. If, however, the passing score resulted in twice the failure rate for blacks as for whites, then regardless of the actual validity of the test for both groups, the passing score would probably be strongly challenged. Prior awareness of the liklihood of these problems can permit collection of relevant data for avoiding or meeting such challenges. A passing score which is unenforceable and unworkable is not going to do anyone much good.

Explainability condemns the jargonistic, pseudo-statistical mumbo-jumbo which so often surrounds the setting of passing scores. If it is not possible to explain to the examinees, their families, the teachers, and the school board how the passing score was set and why, in at least a general way, the system is likely to be unworkable in any situation short of a dictatorship. For this reason, elegant, complex statistical systems are to be avoided unless they offer clear advantages with respect to other categories of concern.

Additional concerns which belong roughly to the explainability category center on the openness of the process. If interested parties are informed and are permitted to provide input to the decision in a meaningful way, the decision system is more likely to be successful.

Taken together, the decision categories attempt to assure that the system embodied in the passing score will be workable. Some of the concerns address the process by which the score is determined, others the process of applying and monitoring its use. Some of the concerns, for example ethical ones, are appropriate regardless of the level and effect of the passing score. Others, for example political ones, may be triggered only when the passing score is set above or below a particular level. The main point to be made is that even when the concern is not directly related to choosing that number which will determine whether the examinee passes or fails, it may end up determining whether the label "pass" or "fail" has the effect of a decision. The letter on the test paper without the decision system to act on it, is meaningless.

Although it has usually been assumed that setting passing scores was a job for educators, this is not necessarily the case. Where the decision system has impacts outside the educational system, it is appropriate to involve those who are affected by the results. It is also desirable to include individuals who have expertise regarding any of the categories of concern, either as participants in the process or as consultants. Teachers "represent important sources of expert judgment on what is taught in the schools, the typical performance levels ... and what can reasonably be expected ..." (Jaeger, 1978, p. 12-13.), and for that reason should normally be included in the process. Experts in curriculum and in testing or psychometrics are also usually valuable contributors to the process. Depending on the nature of the judgment to be made, administrators, future employers of examinees, parents, politicians, university representatives, business leaders, political activists, school counselors, and taxpayers might all have an interest in the decision. Where representatives of these various groups cannot be included as decision-makers,

provision should be made for their input as consultants or in open meetings.

Models of the decision-making process range from the relatively dictatorial and it's twin the "doctor's orders" or professional judgment approach, to the referendum ("let's everybody vote on it"). Because of the number and complexity of concerns, and the need for input from various sources, it is unlikely that a single individual no matter how competent should be the decision-maker. This is not to say that the individual might not reach the same decision as the larger group. Rather, it emphasizes the need to establish a larger base of support for the decision system within the decision-making process.

Also because of the complexity of the information to be considered, a referendum model is inappropriate. It is unlikely that a large group of "voters" could be adequately informed in any financially feasible way. This leaves the various panel approaches which have been recommended. The panel permits representation of various fields of expertise and interested parties without being prohibitively expensive. Shepard (1976) and Jaeger (1978) recommend models in which there is no attempt to have the panel reach consensus. The reasoning is that a "forced" consensus results in artificial agreement. Our society runs on forced consensus, otherwise known as politics. The consensus model, which usually involves confrontation, is messier, noisier, and more difficult to control than the alternatives, and there may be a real concern about whether any decision can be reached. However the alternatives, which usually involve selecting from among the recommended passing scores according to a set criterion (the mean, lowest, highest, etc.) involve substituting the model-maker's judgment for that of the decision-making group. This should be a last-ditch effort akin to submitting a dispute to arbitration, not the usual procedure. The effort to reach consensus requires a level of participation and emotional concern which requires the panel members to share fully in the determination process, and to thereby take responsibility for the score selected. Without that "ownership", they are unlikely to fight to make the system work. It may be tough for the panel to reach a consensus, but it is still

tougher to try to implement a passing score which was never agreed upon in the first place.

Having determined who should be involved in the decision-making process, it is necessary to determine the nature of the data on which the decision will be based. Since the passing score is tied to the instrument on which it is set, test scores will obviously be considered. It seems reasonable that those scores should be from individuals like those for whom the passing score is to be set; either the actual group for which the decision is to be applied or a similar one. The determination of the reliability of the instrument and decision may or may not require a second administration of the test or a parallel form of it to the same or similar examinees. If the test is homogeneous, if no future use of a parallel form with the same passing score is anticipated, or if the decision is being applied only to the group being tested in the process of determining the passing score (as, for example, with a final examination for a single class), then additional test scores may not be needed. If, however, a more complex system is being set up with multiple forms of the test for use with multiple groups of students, it is appropriate to test with at least two forms of the test.

Validity and accuracy concerns require a second type of data, that relating to the competence being assessed. Normally, this will be independent evidence of the examinee's performance. Teacher grades or ratings, the results of other examinations, examinee self-ratings, and other indicators may be appropriate indicators. Where possible, the use of two or more independent indicators of the level of competence is desirable. The indicators should be selected on the basis of their relevance to the decision to be made. If selection is the aim, actual performance in the program for which selection is being made would be appropriate. If certification is being made, independent assessment of the competencies being certified might be made. The function of these "criterion measures" is to help to ground the test score in reality; to assure that it is a valid indicator of the competence being assessed.

Domain-referenced approaches to assessment have a somewhat different approach to "criterion measures". They assume that the actual performance on the items together with a proper description of the nature of the items is sufficient to assure the validity of the score. In this type of system, the independent assessment is generally applied to the items rather than to the examinees. Items are evaluated to determine whether they do, in the opinion of experts, relate to the desired competency. This approach is appropriate in certification systems, but is insufficient for systems where future performance must be predicted. This is because it does not provide any information as to how completely the test has sampled competencies related to future success. If more than one form of the test is to be used based on the same item domain, it is desirable to compare the scores of the same sample of examinees on at least two forms. This is because even where each item is independently considered to be related to the competency, as measured by expert judgment, the item domain may be heterogeneous enough that individuals who do well on one form of the test may do poorly on another. Each item may be a valid indicator of the competence, but the test form as a whole may not be reliable enough to assure that the test score is a valid indicator of domain competence.

Experts are divided on whether the passing point should be determined with respect to the items, the observed test score, the item domain, or some criterion measure. Methods which look directly at the items include the Nedelsky technique and Angoff's adaptation of it, and Ebel's approach (Nedelsky, 1954; Angoff, 1971; Ebel, 1972). In the Nedelsky approach, a qualified judge envisions a student at a given level of competence (for example, the borderline between a D and F course grade). He then strikes from multiple choice items those distractors which this student could be expected to identify as incorrect. The adjusted guessing score, computed over the items on the test (and judges, if more than one is used) yields the numerical passing score. Angoff has judges estimate the probability of the hypothetical student getting the item correct by chance directly, and thereby extends the technique to items not in multiple-choice format. Ebel adds

estimation of the importance of the item, reasoning that if the item is considered critical to the competency, a higher standard should be applied. Jaeger (1978) suggests an alternative for use in high school minimum competency testing, where judges are asked to determine "if a student cannot answer this item correctly, should he or she be denied a high school diploma?" The passing score is the median number of items judged essential, computed across judged.

The advantage of methods applied to items directly is that presumably the information could be added to an item bank and new forms of the test could be generated complete with passing scores. This, of course, ignores the possibility of interactions among items, whereby the presence of one item changes the probability of answering another one correctly. In addition, the proposed methods are weak in that they depend on the implicit performance information held by each judge. As Shepard (1976) points out, expert judgment is based on normative data. Each judge has incomplete, imperfect and ideasyncratic experience by which to judge the expected performance of students with given characteristics. Implicit performance data should not form the basis of passing score decisions except where explicit data is unavailable. Although there is no theoretical reason why item-based methods could not be worked out using explicit performance data, this does not appear to have been done.

Considerable work on setting passing scores to assure some given level of performance on an item domain has been undertaken, primarily by those interested in minimum competency testing. Work has centered on the estimation of a domain score (or its dichotomous counterpart, "mastery" or "non-mastery" status) from the observed test score. Consideration has been given to estimation of reliability (Hambleton & Novick, 1973; Huynh, 1976; Subkoviak, 1978), to the likelihood of making classification errors (Wilcox, 1977), and to the number of items needed for a reliable decision (Millman,1973), but not as to how the passing score on the domain itself should be set. As Jaeger points out, such efforts generally begin with the assumption of fixed categories of "master" and "non-master" as established by some outside source. Wilcox allows the possibility of identifying the categories with

exemplars, but that does not solve the problem of identifying them.

Setting the passing score on a criterion variable is much the same as setting it on an item domain. A good deal has been written about the estimation of "mastery" "non-mastery" status, usually using a regression model, but the determination of the criterion level associated with "mastery" has not been considered. The procedure of setting the passing score with respect to the item domain or a criterion variable (or, for that mater, a latent trait) simply defers the judgment process to that reference domain. Then a second step, that of estimating whether the reference "mastery" level has been reached on the basis of the observed score, is required. Unless there is clear evidence for the appropriateness of setting the passing score on the basis of some reference domain (for example in a selections system where success or failure in the program for which the selection is being made is easily observed and clearly defensible), the passing score should be set on the observed score distribution. To do otherwise is to increase the difficulty of explaining what was done and why, to decrease the clarity of meaning associated with "passing", and to increase the importance of errors of estimation in the system of decision-making.

Passing scores are most commonly set on the basis of the observed test score distribution, examined in conjunction with collateral information as to the competence of the examinees. Such data may be collected and used explicitly (for example by comparing the test scores of examinees with their course grades). It may also be used implicitly. Two common methods of setting passing scores by using implicit criterion information are the Borderline Group Method and the Contrasting Group Method, both of which involve the use of exemplars (Popham, 1978). The Borderline Group Method involves selecting some specific point in the score distribution of students identified as being on the "borderline" between competence and incompetence. Commonly the mean or median would be selected. In the Contrasting Group Method, the score distributions of examinees judged to be clearly competent and clearly incompetent are examined and

some point relating to the area of intersection selected as the passing point. Commonly the midpoint of the intersection, the actual score at which the two distributions intersect, the score below which no clearly competent student scored, or the score above which no clearly incompetent student scored would be selected. These methods are weak in that they do not make explicit the means by which the exemplars were selected, thus disguising the nature of the criterion variables. In addition, by treating the groups as dichotomous ("competent or incompetent", "borderline or not borderline"), they loose a good deal of information, both with respect to the level of criterion performance and with respect to the complex set of criterion measures which may have been used in arriving at the classification. Use of the actual test scores and criterion measures allows for the explicit use of the data and leaves open the possibility of dichotomizing it in the analysis should that be desirable without eliminating the possibility of using the full distribution of scores.

In addition to the test scores (on one or more forms) and collateral criterion information for a representative group of examinees, the passing score decision process will often require information on groups of examinees who are not representative. Some of this information may be collected by identifying the individuals as they occur in the regular sampling process. For example, data collected for a sample of 12th grade students might be identified as belonging to individuals of particular ethnic groups, with non-English language backgrounds, or having specific handicapping conditions. Where the incidence of such individuals in the regular population sampled is low, however, the data may need to be collected for additional individuals in the special populations.

An additional special population with respect to competence certification programs is the individuals who are placed in a remediation program. It is of concern to know the likelihood of these examinees passing the test after the remedial instruction. As a control measure, it may also be a good idea to find out whether individuals who fail the test on the first round are likely to be able to pass it on the

second without additional formal instruction, either on the basis of their own review efforts or of increased familiarity with the test, less nervousness, or whatever.

While concern for the characteristics (competency level) of those who pass or fail the test forms a central concern for the passing score system, so is concern for the numbers of examinees placed in each category. It is the "how many" which will often determine the impact of the decision system on other social systems. A careful sampling of the relevant examinee population will permit estimates of the numbers of individuals passing and failing given any particular passing score. If more than one competency is being assessed, however, it may be important to know whether the same or different individuals are likely to fail different competencies. If, for example, three competencies are required for high school graduation with those failing any competency being retained for remedial work, and 4% of the student are expected to fail each of the tests, it becomes of interest to know whether the school must plan for 4% or 12% of the students to be retained.

The data, therefore, consists of test scores and collateral competence information for a sample of the examinee population and additional samples of special examinee populations of interest (remedial students, the handicapped, ethnic groups, etc.). The test scores are based on two or more forms of the test, if more than one form is planned. Data is collected at both the item and test level. Three types of data analysis are conducted: analysis of the test, analysis of the characteristics of examinees with specific test scores, and analysis of the numbers of examinees classified (or misclassified) as "passing" or "failing" given tentative passing points.

Test analysis is covered in most books on psychometrics, and will probably be conducted by a psychometrician or statistician associated with the passing score effort. In addition to the estimation of test reliability and validity (with respect to the various conditions of testing and the various criterion measures, respectively), the analysis should include item analysis information relating correct item responses to the

competence criterion measures. Tables 1 and 2 suggest presentations for regular and remedial students treated categorically. Examination of the data suggests a problem with item 3: students judged competent and incompetent performed about equally well on the item, and remedial work did not improve student performance on it. In such a case, the item should be examined to determine whether it is flawed in some way, or whether the instructional system may have failed to cover the material which it assesses. If many flawed items are found on the test, the test itself may have to be revised or at least rescored before further analysis will be meaningful.

Once the adequacy of the test items has been assured, analysis should characterize the relationship between the test score and the various criterion information. Figure 1 presents a possible graphic presentation for two forms of the test for the regular and remedial populations. Continuous information might be presented in the form of scattergrams or regression equations (with confidence bands) if the decision-makers are able to understand the information. A key concern is to keep data presentation as clear and simple as possible, since the individuals who need to understand it will not, in general, have been trained in the field.

The third type of analysis is based on the numbers of students who can be expected to pass or fail given any particular passing score. Although it will be easier to compare the passing rates for two or more forms of the test if this information is presented in percent passing form, the impact on the system is more clearly seen in the actual number of students. Alternate forms may therefore be scaled to some single number of students (for example, the number expected to take the test next year) to give estimated number of students failing. To avoid confusion, it is desirable to present either "number passing" or "number failing" data, not both. Either can be generated from the other, given the total N, and generally the major concern is with one group or the other (in a certification system, only the "failing" group needs to be retained; in a selection system, only those selected become involved in the program). Data should, of course, be presented in cumulative form. Table 3 and Figure 2 present possible tabular and graphic presentations.

Finally, data relating competence and numbers of students failing jointly to the test score can be presented to give a more complete, if more complex, picture of the situation. It may be desirable, because of the complexity of the information, to have a consultant explore the data to determine the most appropriate means for categorizing it. The presentation offered here includes only one competence measured in three categories (not competent, borderline, and competent) for two populations (regular and remedial) on two test forms (A and B). The actual data might well involve several continuous criterion measures for many populations on many test forms covering several different competencies. Table 4 presents one possible tabular format.

Although some of the data analysis may be conducted prior to consideration of passing scores (for example, item analysis), most of it should be handled interactively with the analysis redirected to answering questions formed by the judges. The process aims at providing the judges with the clearest possible picture of the test data. In addition, consideration by the judges of the categories of concern might well dictate the collection of additional information unrelated to the actual test scores. The cost of remediation, the degree of community satisfaction with the competence of past graduates, the failure rate in college or business and the incompetencies to which it is attributed, the anticipated cost of the testing program, the number of remedial teachers available, and other considerations may require the collection of data. For large and complex decision systems, a regular research and evaluation effort may be needed. For smaller and more direct situations, it may simply be a matter of consulting with the teacher of the course to which passing students will be routed to determine the degree of satisfaction with the preparation of previous students. The same concerns apply, they are simply scaled down.

With the judges selected, the data collected, and data analysis begun, the process of making judgments based on the data can be initiated. This is the heart of the process. Whether it is undertaken by one or two instructors of the same course, or by a panel of community representatives

charged with making decisions for the entire educational system, it may not be taken lightly. The decisions made will affect peoples' lives and their livlihoods, their self-esteem, and their future successes. Although the system to be described is a complex and relatively formal one, designed to meet the needs of a decision system which will affect many people, smaller, less formal systems will share many of its features.

The first meeting of the judges has as its aim to introduce the individuals to one another and to the task at hand. It is appropriate, if the individuals are not already acquainted, to make introductions (perhaps providing each with a set of names and addresses and brief resumes). The framework within which the judges must function should be laid out. It is appropriate to include the nature of their task (which competencies or tests they are responsible for, and which populations of examinees), the lattitude they have with respect to other social systems (can they call on consultants? recommend changes in remedial programs?), and their time-table. One or more representatives of the major system affected (the public school, the company for which they are to select employees) should present an overview of the current status of the organization, its clientele, and its plans and hopes for the immediate future. A description of the current political and social pressures on the organization, both within and from the outside, and of the current financial status should be included. If the judges represent particular aspects of this situation, they may themselves make some of the presentations. Finally, the judges should become aquainted with the examination and its consequences. Copies of the test should be distributed so that they may examine the actual items. Details of the expected activities of those who pass and fail the test should be presented as completely as possible. While much of this information will be examined in greater detail later, the general overview permits the judges to orient themselves to the task at hand and to its context.

Jaeger suggests that the judges actually take the test. Although there is no better way to learn about test items than to have to respond to them, that approach has several draw-backs.

For one thing, it can be extremely time-consuming. It can also make the judges nervous and anxious about their own performance, a concern which is generally irrelevant to setting the passing score since they are not examinees in the decision system. If it is considered desirable to ask judges to respond to items, a sample of items may be used to both shorten the time required and prevent the judges from comparing their scores with the actual examinee distribution.

In most cases after the first meeting the judges will set their own schedule and agenda, allowing time for presentations by their own members and consultants, for public input, and for learning to understand data-based presentations and deliberating on them. They will be responsible for directing their staff as to what questions they wish answered by collecting and analyzing data, what records are to be kept of their meetings, and when and how their meetings are to be conducted. In arranging to inform themselves, the judges should be aware that they will be able to attend to only a limited amount of information at once. This is particularly true of test data, especially for those unused to dealing with it. Too much data presented in a single session can result in confusion and a tendency to ignore what may be important information. The desire to restrict the data presented must be counter-balanced, however, by the need to present a rather complete picture rather than a single facet of it. If only gloomy data on impact, for example, is presented, the judges may tentatively set the passing score much too low only to have to change it drastically when the more encouraging remediation information is added.

Not only is setting tentative passing scores which are way out of line a waste of time, it may also tend to set judges' minds on a single number too early in the process. It is important, given the wealth of information to be considered, that the judges remain flexible and open-minded as long as possible. For this reason, it is best not to single out passing scores early in the process (even by, for example, having each judge suggest a passing score and then inspecting the distribution). Ownership of particular scores, with

the tendency to defend them, should be avoided. Rather, the emphasis may be placed on narrowing the range within which the passing score will be eventually placed. To facilitate this process, and prevent jumptin too fast for a single score, the initial data presentation should be at full scale, including the full range of test and criterion scores, and using 5 or 10 point intervals on the axes of figures or as the categories for tables. The first decision is then to eliminate one or both tails of the distribution from consideration. This may be a simple matter of excluding those scores lower or higher than anyone achieved, or those which would result in everyone passing or everyone failing. It might even go so far as to eliminate passing scores which would result in more than 60 or 70 or 80 percent of the examinee population failing.

Once the first range restriction has been accomplished, hopefully giving everyone the sense that joint decisions are indeed possible, the data may be rescaled to the new range. Again the discussion centers on narrowing the range, using the now more specific intervals (perhaps 3 to 5 score point widths) available. As each range restriction is accomplished, the data is rescaled and presented more specifically. At any point where the group cannot agree on a further range restriction, it should determine what data would be of assistence in doing so, and arrange for its collection and presentation. Figure 3 suggests the additional level of detail possible with rescaling.

If the number of judges is large, the discussion of possible range restriction may be done in groups of 3 to 6 individuals, changing the group membership on each round. This encourages the judges to become personally acquainted, communicates the individual judge's expertise more widely, and helps to prevent the formation of cliques.

Once the range is restricted to fewer than five points, the judges may wish to consider the appropriateness of selecting a lower passing score with the intention of raising it in the next year or two, as the system adjusts to the new procedures. In any case, the possibility of adjusting the passing score should always be available, should a change in the situation dictate.

The consideration and reconsideration of data is likely to be tiring, and should usually be spaced out over several meetings. Varying the type and format of data presentations, and interspersing information about curriculum content and financial considerations with test data may be helpful. If the judges are responsible for passing scores for more than one test, moving from one to another in the same meeting may provide a change of pace, especially if the considerations are at different stages. However, the deliberations on one competency test should not be completed individually if the desired result is a multi-competency system, since like other systems they can be expected to interact. Joint failure rates for the several competencies are important considerations, and should be included before any individual passing scores are determined. Table 5 presents a possible format for joint failure rate data. It should be noted that where joint failure rates are a concern, joint remediation rates should be also. It is important to know whether remedial efforts for those who are deficient in several competencies are likely to help the same individuals regardless of the specific competency, or to help different individuals in different competencies (leaving them in the remedial program for the remaining competency).

Once the passing score(s) are agreed on, the program can be built around them. It should provide for the details of test administration and scoring, score reporting, an appeals system, and continuous monitoring and evaluation of the results. Those aspects of the decision system which relate to the areas of concern should be given careful attention to assure that the final product is as workable as possible. The judges or their delegated representatives must take continuing responsibility for noting shifts in score distributions for the same or new forms of the test, for determining whether new educational programs or new competencies associated with existing programs require changes in the program, and for seeing that the program is implemented in ethical, efficient, and educationally beneficial ways.

The passing score is a test score which represents more than items correct. It represents a catagorization which not only describes individuals, but guides their futures. It is an expression of expectations and hope both for the system and and for the person. It deserves the most thoughtful, the most careful and sensitive judgment we can provide.

Table 1
Tabular presentation of
item data

Item #	Percent of students responding correctly			
	Total (N=300)	Competent (N=191)	Borderline (N=78)	Not Competent (N=31)
1	52	86	45	23
2	73	97	60	54
3	40	42	39	37
4	67	75	64	21
5	60	89	52	47

Table 2
Tabular presentation of
remediation data

Item #	Percent of students responding correctly		
	Before remedial work	After remedial work	Difference
1	24	46	22
2	56	82	26
3	30	28	-02
4	22	56	34
5	48	80	32

Table 3
Number of students failing

Passing score set at	Regular students		Remedial students	
	Form A (N=300)	Form B (N=298)	Form A (N=27)	Form B (N=29)
95	293	284	26	29
90	250	277	23	25
85	223	254	20	21
80	189	211	17	19
75	165	183	13	15
70	142	158	10	12
65	102	111	8	7
60	72	80	6	5
55	41	39	4	3
50	22	19	4	2

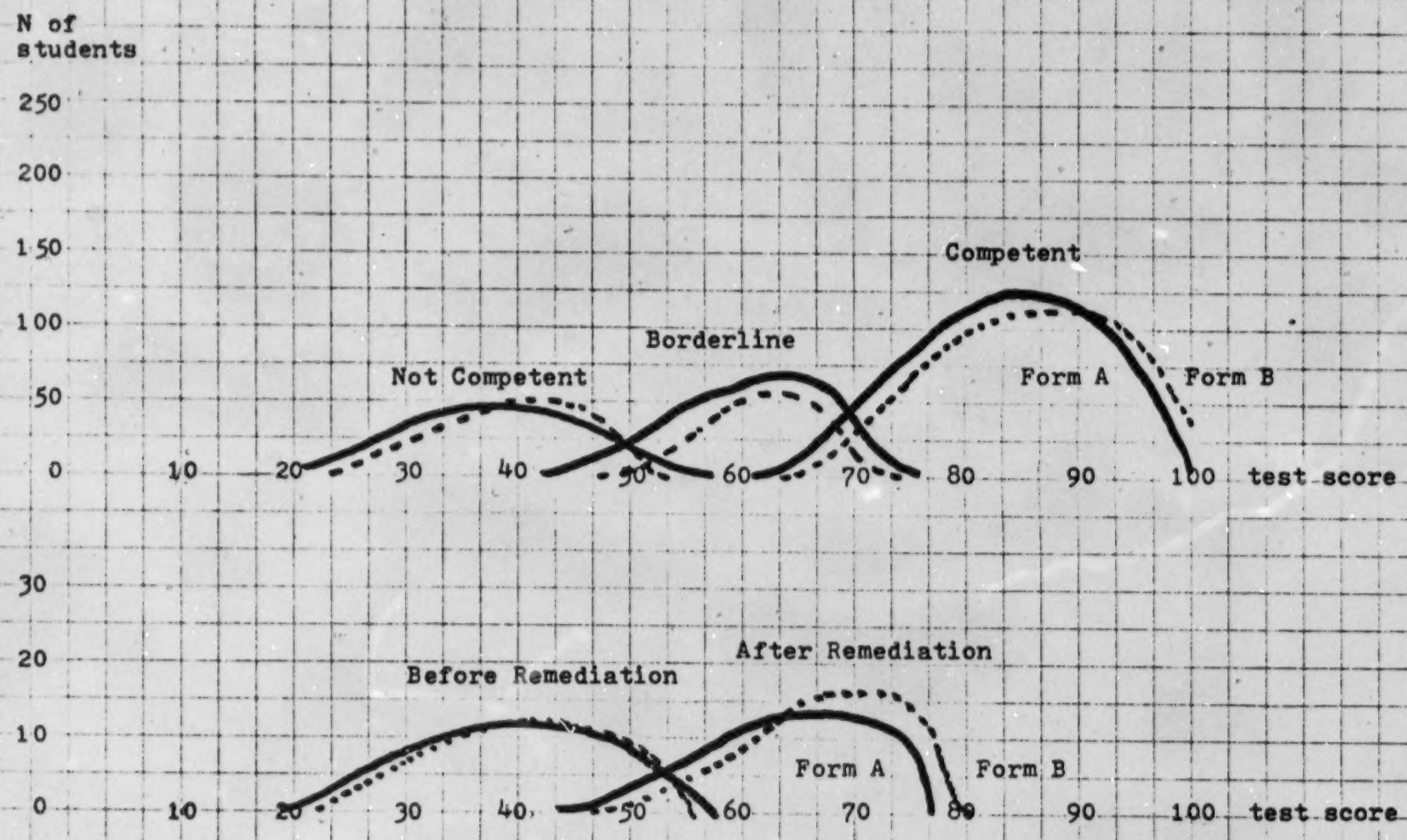
Table 4
Cumulative number of students failing

Passing score set at	Competent (N=50)		Borderline (N=50)		Not Competent (N=50)		Post Remedial (N=50)	
Form:	A	B	A	B	A	B	A	B
95	35	37	42	41	50	48	42	45
90	27	31	39	39	47	45	38	40
85	15	20	32	36	45	44	33	35
80	9	11	28	30	41	37	25	30
75	5	6	18	17	34	29	17	21
70	2	1	8	9	26	20	12	15
65	0	0	2	3	15	13	7	8
60	0	0	1	1	7	5	3	3
55	0	0	0	1	2	1	3	2

Table 5
Joint failure rate for math and reading

Passing score		68	69	70	71	72
	Individual failure rate	2%	3%	5%	7%	8%
65	3%	4%	5%	7%	8%	9%
66	4%	5%	6%	8%	10%	11%
67	6%	7%	8%	9%	11%	12%
68	9%	10%	11%	12%	14%	15%
69	11%	11%	12%	13%	16%	17%

Figure 1
Student test performance by category
and test form



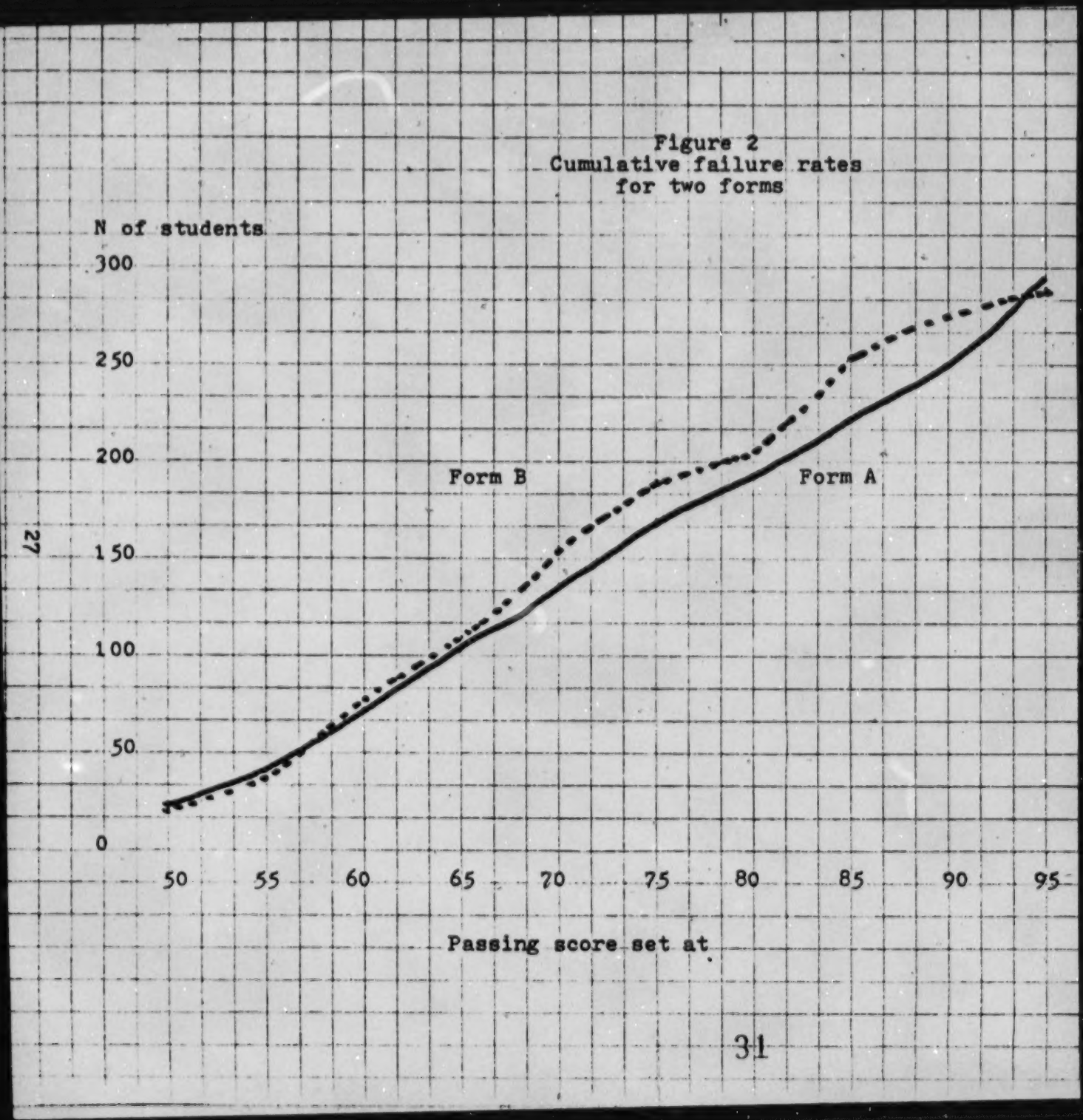
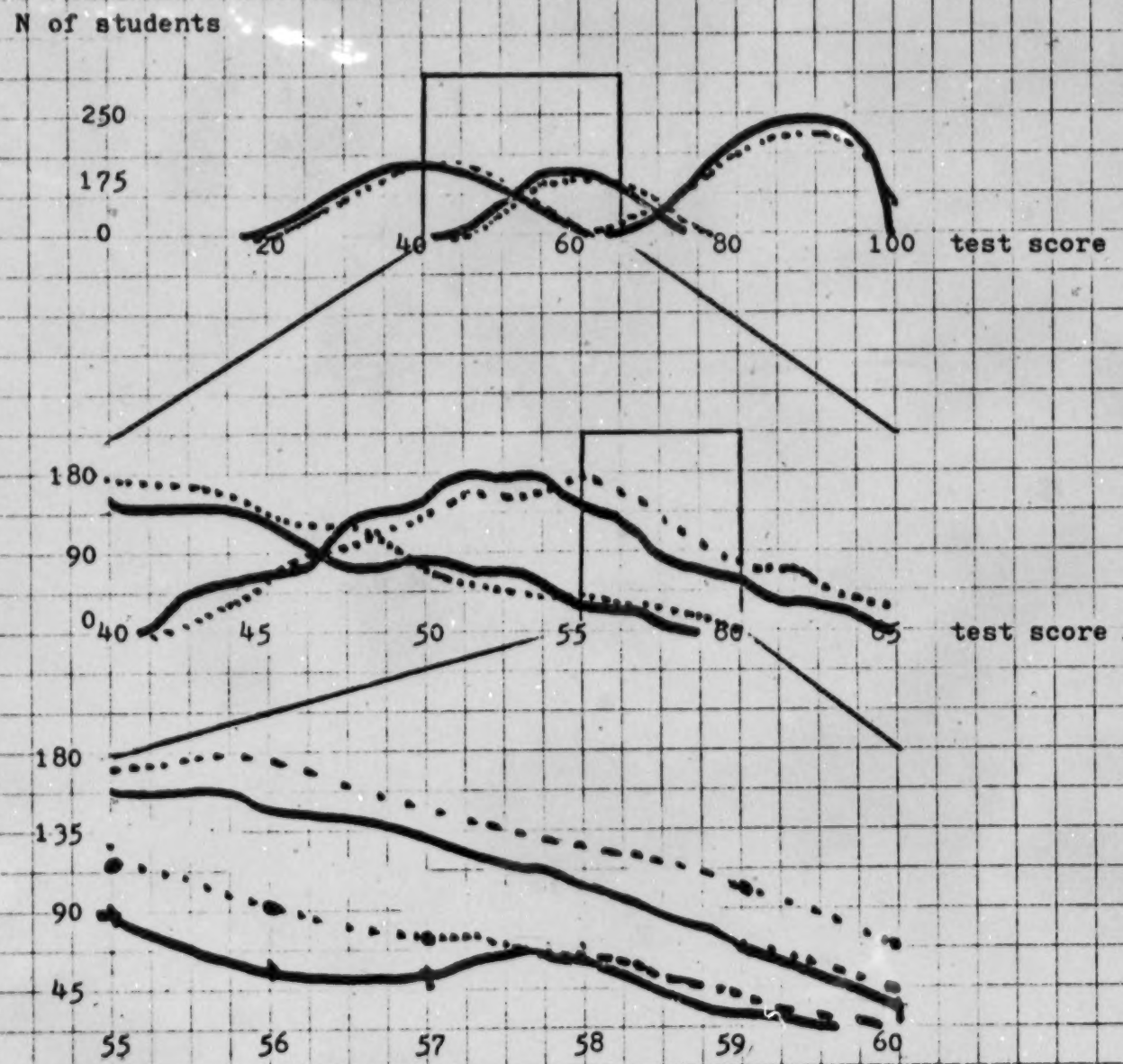


Figure 2
Cumulative failure rates
for two forms
N of students
300
250
200
150
100
50
0
Form B
Form A
50
55
60
65
70
75
80
85
90
95
Passing score set at

Figure 3

Test score distributions
at different scales



Reference Notes

1. Jaeger, R. M. A proposal for setting a standard on the North Carolina high school competency test. Paper presented at the meeting of the North Carolina Association for Research in Education, Chapel Hill, North Carolina, 1978.
2. Linn, R. L. Criterion-referenced testing, book in preparation, 1977.

References

American Friends Service Committee. A citizen's introduction to minimum competency programs for students. Jackson, Miss.: Southeastern Public Education Program, 1978. (ERIC Document Reproduction Service No. ED 155 200)

Anastasi, A. Psychological testing, 4th ed. New York: Macmillan Publishing Co., Inc. 1976.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement, Washington, D. C.: American Council on Education, 1971.

Anderson, B. B. & Lesser, P. The costs of legislated minimum competency requirements, Phi Delta Kappa? 1978, 59, 606-608.

Berk, R. A. Determination of optinam cutting scores in criterion-referenced measurement. The Journal of Experimental Education, 1976, 45, 4-9.

Burton, N. W. Societal standards. Journal of Educational Measurement, 1978, 15, 263-271.

Brickell, H. M. Seven key notes on minimum competency testing. Phi Delta Kappa? 1978, 59, 589-592.

Conaway, L. E. Discussant comments: setting performance standards based on limited research. Florida Journal of Educational Research, 1976, 18, 35-36.

Ebel, R. L. The case for minimum competency testing. Phi Delta Kappan, 1978, 59, 546-549.

Ebel, R. L. Essentials of educational measurement. New Jersey: Prentice Hall, Inc., 1972.

Elford, G. and others. Handbook of management standards for awarding high school diplomas. Massachusetts: Massachusetts Advisory Council on Education, 1975. (ERIC Document Reproduction Service No. ED 147 311)

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Fisher, T. H. Florida's approach to competency testing. Phi Delta Kappan, 1978, 59, 599-602.

Fremer, J. & Dwyer, C. A. Setting standards for basic skills reading assessment. Princeton, N. J. : Educational Testing Service, 1977. (ERIC Document Reproduction Service No. ED 148 833)

Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.

Hambleton, R. K. & Novick, M. P. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hart, G. K. The California pupil proficiency law as viewed by its author. Phi Delta Kappan, 1978, 59, 592-595.

Huynh, H. On the reliability of decisions in domain referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Jaeger, R. M. Measurement consequences of selected standard-setting models, Florida Journal of Educational Research, 1976, 18, 22-27.

Levin, H.M. Educational performance standards: image or substance? Journal of Educational Measurement, 1978, 15, 309-319.

McClung, M. S. Are competency testing programs fair? legal? Phi Dalta Kappan, 1977, 59, 397-399.

Meskauskas, J.A. Evaluation models for criterion-referenced testing: views regarding mastery and standard-setting. Review of Educational Research, 1976, 46, 133-158.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Pipho, C. Minimum competency testing in 1978: a look at state standards. Phi Delta Kappan, 1978, 59, 585-587.

Popham, W. J. Setting performance Standards. Los Angeles: Instructional Objectives Exchange, 1978.

Popham, W. J. & Carlson, D. Deep dark deficits of the adversary evaluation model. Educational Researcher, 1977, 6, 3-6.

Quirk, T. J. Some measurement issues in competency-based teacher education. Phi Delta Kappan, 1974, LV, 316-319.

Scriven, M. How to anchor standards. Journal of Educational Measurement, 1978, 15, 273-275.

Shepard, L. A. Setting standards and living with them. Florida Journal of Educational Research, 1976, 18, 23-32.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15, 111-116.

Wilcox, R. Estimating the liklihood of false positive and false negative decisions in mastery testing: an empirical bayes approach. Journal of Educational Statistics, 1977, 2, 289-307.

Wise, A. E. Minimum competency testing: another case of hyper-rationalization. Phi Delta Kappan, 1978, 59, 596-598.